2η ΕΡΓΑΣΙΑ ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΠΛΑΤΩΝ ΚΑΡΑΓΕΩΡΓΗΣ 3180068

ΝΙΚΟΛΕΤΑ-ΚΛΕΙΩ ΠΑΤΑΤΣΗ 3180266

1Ο Ερώτημα

Α) Με βάση τα φροντιστήρια και τις διαλέξεις του μαθήματος, αρχικά πρέπει να ελέγξουμε τις 3 συνθήκες προκειμένου να συμπεράνουμε, εάν τα δεδομένα μας είναι κατάλληλα, για να εφαρμόσουμε τις μεθόδους της συμπερασματολογίας που γνωρίζουμε. Αρχικά, σύμφωνα με την εκφώνηση τα δεδομένα μας προέρχονται από τυχαία δειγματοληψία, επομένως η πρώτη συνθήκη τηρείται. Δεύτερον, έχουμε πάνω από 15 τιμές, συγκεκριμένα έχουμε 20, άρα μπορούμε να πούμε με σχετική ασφάλεια, ότι τηρείται και η δεύτερη συνθήκη, καθώς το σχήμα μας δεν παρουσιάζει σημαντικές αποκλίσεις. Τέλος, έχουμε outliers και μένει να αποφασίσουμε, κατά πόσο επηρεάζουν σημαντικά το διάστημα. Θεωρούμε ότι δεν το επηρεάζουν τόσο, ώστε να μην μπορούμε να εφαρμόσουμε τις μεθόδους της συμπερασματολογίας.

Β) Αρχικά, με βάση το φροντιστήριο πρέπει να χρησιμοποιήσουμε τον τύπο «mean(x) + c(-1,1)*t*sd(x)/sqrt(n)». Από όλα τα στοιχεία του τύπου το μόνο που χρειάζεται περαιτέρω ανάλυση είναι το t, όπου είναι η τιμή της κατανομής και για να υπολογιστεί χρειάζεται ένα ποσοστημόριο και ο βαθμός ελευθερίας όπου είναι n-1, με n των αριθμό των δεδομένων. Δηλαδή, «t<- -qt(0.025, df=n-1)» στο t*. Το 0.025 το βρίσκουμε με την εξής λογική. Ψάχνουμε 95% πιθανότητα επομένως από 0 μέχρι 1 είναι 0.95 και άρα έξω από το ζητούμενο εμβαδόν έχουμε 0.05 πιθανότητα. Όμως η κανονική κατανομή είναι συμμετρική, επομένως το ποσοστημόριο είναι 0.05 διά 2 άρα 0.025. Εν κατακλείδι, το διάστημα εμπιστοσύνης είναι 51.41365 103.38635.

2Ο Ερώτημα

Α) Το λάθος εδώ είναι ότι ο τύπος τυπικής απόκλισης δειγματικού μέσου είναι σ/sqrt(n) και όχι σ/n.

Β) Το λάθος στην περίπτωση αυτή είναι ότι με βάση τη θεωρία το Η0: μ=μ0 απαιτεί το μ0 να είναι σταθερή και γνωστή τιμή και στο παράδειγμα εδώ δίνεται ότι μ= $\bar{x}$, και το $\bar{x}$ = 10 άρα μ0= $\bar{x}$. Αυτό προφανώς δεν ισχύει, γιατί το $\bar{x}$ είναι ο εκτιμητής και το μ0 είναι μία σταθερή τιμή που σημαίνει ότι δεν μπορεί να γίνει τέτοια αντικατάσταση στον τύπο του Η0. Επίσης, οι υποθέσεις αφορούν μόνο παραμέτρους του πληθυσμού και δεν εξαρτώνται από το δείγμα.

Γ) Ανατρέχουμε στην R και κάνουμε αντικατάσταση στους τύπους. Θεωρούμε τυχαίο σύνολο n=15 στοιχείων (43,45,45,45,45,45,45,45,45,45,45,45,45,45,47) όπου έχει xbar=45, sd=0.7559289 και από την εκφώνηση έχουμε μ0=54. Τα δεδομένα αυτά είναι τυχαία, έχουμε n>=15 και δεν παρατηρείται κάποιο outlier, επομένως μπορούμε να χρησιμοποιήσουμε τις μεθόδους συμπερασματολογίας που έχουμε διδαχθεί. Υπολογίζω το z = (xbar-μ0)/(σ/sqrt(n)) το οποίο, έχει τιμή περίπου -46.1. Έπειτα, αφού έχω εναλλακτική περίπτωση Ηα: μ0>54, εφαρμόζω τον τύπο 1-Φ(z). Ο τύπος αυτός δίνει ένα pvalue που αγγίζει το 100%. Το βέβαιο είναι, ότι το pvalue είναι μεγαλύτερο από ένα αυθαίρετο όριο α=5%, γεγονός που μας οδηγεί στο συμπέρασμα ότι δεν μπορεί να απορριφθεί η Η0. Άρα η πρόταση είναι λάθος.

Δ) Για να απορρίψουμε τη μηδενική υπόθεση, πρέπει το pvalue<α, όπου α ένα αυθαίρετο όριο, που συνήθως ορίζεται ως α=5% ή α=10% ή α=1%. Εδώ pvalue=0.52, που σημαίνει ότι είναι πολύ μεγαλύτερο από το όριο α. Αυτό μας αποτρέπει από το να απορρίψουμε τη μηδενική υπόθεση, χωρίς να σημαίνει όμως και ότι την αποδεχόμαστε. Όπως και να χει, η πρόταση είναι λάθος, αφού δεν απορρίπτουμε τη μηδενική υπόθεση.

3ο Ερώτημα

Α) Ο τύπος που χρειαζόμαστε σε αυτή την περίπτωση είναι 1-Φ(z) . Τρέχουμε στην R την εντολή 1-pnorm(z). Το αποτέλεσμα που μας δίνει είναι 0.09012267.

Β) Ο τύπος που χρειαζόμαστε σε αυτή την περίπτωση είναι Φ(z) . Τρέχουμε στην R την εντολή pnorm(z). Το αποτέλεσμα που μας δίνει είναι 0.9098773.

Γ) Ο τύπος που χρειαζόμαστε σε αυτή την περίπτωση είναι 2Φ(-|z|) . Τρέχουμε στην R την εντολή 2*pnorm(-abs(z)). Το αποτέλεσμα που μας δίνει είναι 0.1802453.

4ο Ερώτημα

Α) Αφού το διάστημα εμπιστοσύνης είναι 95%, τότε το α=1-C, είναι ίσο με 5%. Άρα α=0.05 και έχω pvalue=0.04 που σημαίνει ότι έχουμε σημαντικές ενδείξεις, που μας οδηγούν στην απόρριψη της μηδενικής υπόθεσης. Αφού απορρίπτεται η μηδενική υπόθεση, τότε καταλήγουμε ότι το μ0=30, δεν θα βρίσκεται στο 95% διάστημα εμπιστοσύνης. Συγκεκριμένα, όποτε η μηδενική υπόθεση απορρίπτεται, τότε το μ0 ή αλλιώς η αξία της μηδενικής υπόθεσης, δεν θα βρίσκεται μέσα στο αντίστοιχο ποσοστό εμπιστοσύνης, ενώ ισχύει και το αντίστροφο, δηλαδή εάν δεν ανήκει το μ0 στο ποσοστό εμπιστοσύνης, τότε συνεπάγεται ότι απορρίπτεται η μηδενική υπόθεση. Αυτό ισχύει, διότι το επίπεδο εμπιστοσύνης και το επίπεδο σημαντικότητας ορίζουν η κάθε μία, μία ξεχωριστή απόσταση, οι οποίες εν τέλει ταυτίζονται. Αναλυτικότερα, το επίπεδο εμπιστοσύνης, εκφράζει πόσο κοντά είναι τα όρια εμπιστοσύνης από το δειγματικό μέσο και το επίπεδο σημαντικότητας, την απόσταση του δειγματικού μέσου από την αξία της μηδενικής υπόθεσης, όταν αυτή είναι στατιστικά σημαντική.

Β) Με βάση το Α, από τη στιγμή που δεν ανήκει στο 95% διάστημα εμπιστοσύνης, τότε είναι προφανές ότι δεν ανήκει ούτε στο 90% διάστημα, αφού εξακολουθεί το pvalue να είναι μικρότερο από το α. Επίσης, μπορούμε να το συμπεράνουμε από το σχήμα πιθανοτήτων στις

διαφάνειες. Συγκεκριμένα, έχουμε αποδείξει ότι δεν μπορεί να ανήκει το μ0=30 σε ένα μεγαλύτερο ποσοστό, επομένως όταν αυτό μικραίνει, μικραίνει και το z* και άρα η πιθανότητα να βρίσκεται εκτός του διαστήματος εμπιστοσύνης.

5^{Ο} Ερώτημα

Α) Αρχικά επιλέγουμε να πάρουμε τον πίνακα από το PDF και αφού τον κάνουμε paste σε ένα txt αρχείο, τον κάνουμε import στην R. Έπειτα, θεωρούμε ότι το «6» βάρος στη σειρά 14 είναι τυπογραφικό λάθος και επιλέγουμε να το αγνοήσουμε. Έπειτα, αντιμετωπίσαμε πρόβλημα με τα ελληνικά και επιλέξαμε να τα αλλάξουμε σε αγγλικά, προκειμένου να τα καταλαβαίνει η R και να μην βγάζει το «unexpected input error»(παρά την επιλογή UTF-8 στο import, όταν τρέχαμε εντολές τύπου mean(x$ΒΑΡΟΣ) είχαμε το παραπάνω error). Προτού προβούμε σε οποιονδήποτε υπολογισμό, πραγματοποιούμε μία διερεύνηση για να κρίνουμε εάν τα δεδομένα είναι κατάλληλα προς επεξεργασία. Αφού τα δεδομένα προέρχονται από τυχαία δειγματοληψία, το n> 15 και δεν έχουμε κάποιο outlier καθώς αγνοούμε το «6», καταλήγουμε ότι τα δεδομένα είναι κατάλληλα και μπορούμε να χρησιμοποιήσουμε τις ανάλογες μεθόδους συμπερασματολογίας. Υπολογίσαμε την μέση τιμή και την τυπική απόκλιση με τις εντολές «mean(x$BAROS, na.rm=TRUE), sd(x$BAROS, na.rm=TRUE)» και τις καταχωρήσαμε στις μεταβλητές m,s. Ύστερα, υπολογίσαμε και το t αντίστοιχα με την πρώτη άσκηση με το ποσοστημόριο να είναι 0.025, λόγω του ζητούμενου 95% διαστήματος εμπιστοσύνης και με το βαθμό ελευθερίας 24-1, αφού έχω n=24 τιμές. Εν τέλει, τρέχουμε την εντολή m + c(-1,1)*t*s/sqrt(n) για να πάρουμε ως αποτέλεσμα το διάστημα 69.57826 78.00507.

Β) Εδώ καταλήξαμε ότι είναι πολύ βοηθητική η t.test με two sample test και τρέξαμε την παρακάτω εντολή με το αντίστοιχο αποτέλεσμα:

```
> t.test(x$BAROS[x$FYLLO=="A"], x$BAROS[x$FYLLO=="G"],conf.level=0.8)

        Welch Two Sample t-test

data:  x$BAROS[x$FYLLO == "A"] and x$BAROS[x$FYLLO == "G"]
t = 2.9923, df = 19.005, p-value = 0.007486
alternative hypothesis: true difference in means is not equal to 0
80 percent confidence interval:
  5.948055 15.436561
sample estimates:
mean of x mean of y
 78.69231  68.00000
```

Το αποτέλεσμα όπως βλέπουμε, σύμφωνα με την R είναι 5.948055 15.436561.

Γ) Για να συμπεράνουμε αν οι δύο αυτές μεταβλητές έχουν σχέση θα φτιάξουμε μία υπόθεση όπου H0: μ0=0 όταν είναι ίδιο το βάρος ανεξαρτήτως καπνίσματος και Ha: μ0!=0 όταν το κάπνισμα έχει επιρροή. Αυτό γίνεται εύκολα όπως βλέπουμε παρακάτω:

```
> t.test(x$BAROS[x$KAPNISTHS=="YES"],x$BAROS[x$KAPNISTHS=="NO"])

        Welch Two Sample t-test

data:  x$BAROS[x$KAPNISTHS == "YES"] and x$BAROS[x$KAPNISTHS == "NO"]
t = 1.2597, df = 19.281, p-value = 0.2228
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.403049 13.717334
sample estimates:
mean of x mean of y
 76.80000  71.64286
```

Όπως βλέπουμε το pvalue αγγίζει το 22.3% και επομένως δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση. Ωστόσο, τα mean φαίνεται ότι παρουσιάζουν μια σχετική απόκλιση. Συμπεραίνουμε ότι είναι πιθανό να υπάρχει επιρροή του καπνίσματος, αλλά όχι αρκετή με τα τωρινά δεδομένα για να απορρίψουμε τη μηδενική υπόθεση, αφού υπάρχει 22.3% πιθανότητα ότι θα εμφανιστούν τιμές που θα επιβεβαιώνουν την H0.

6^{Ο} Ερώτημα

Α) Τα δεδομένα προέρχονται σύμφωνα με την εκφώνηση από τυχαίο δείγμα, επομένως αφού έχουμε, n>15 δείγματα και δεν παρατηρούνται outliers που να επηρεάζουν σημαντικά τα δεδομένα τότε συμπεραίνουμε ότι είναι κατάλληλα για τις μεθόδους συμπερασματολογίας που γνωρίζουμε.

Β) Η τυπική απόκλιση σύμφωνα με την R είναι 0.6008766. Αντίστοιχα, η μέση τιμή είναι 5.5.

Γ) Αρχικά επιχειρούμε να «χτίσουμε» τον τύπο «mean(x) + c(-1,1)*t*sd(x)/sqrt(n)» χρησιμοποιώντας τα παραπάνω δεδομένα. Έχουμε έτοιμη τη μέση τιμή και την τυπική απόκλιση από το προηγούμενο ερώτημα επομένως μένει να υπολογίσουμε το t. Να σημειώσουμε ότι προτιμούμε τη χρήση του t από το z λόγω σχετικά μικρού αριθμού δεδομένων. Θέτουμε το t ως «-qt(0.025, df=n-1)». Το αποτέλεσμα είναι το διάστημα [5.218781, 5.78121219].

7^{Ο} Ερώτημα

Θεωρούμε ότι οι 2 εκτιμήσεις, είναι εξαρτημένες η μία από την άλλη. Αυτό διότι και οι δύο εκτιμήσεις, γίνονται πάνω στο ίδιο όχημα στην ίδια κατάσταση, ανεξάρτητα με το αν υπάρχει περίπτωση το συνεργείο να αυξάνει τις δικές του εκτιμήσεις. Αν τις αυξάνει, τις αυξάνει σε μία ρεαλιστική βάση, ανάλογα με την κατάσταση του οχήματος. Δεν μπορεί δηλαδή, να εκτιμήσει ο ξένος εμπειρογνώμονας σε μία τυχαία περίπτωση, κόστος 150 και το συνεργείο 3000. Αντίστοιχα αν δεν τις αυξάνει, τότε πρέπει οι 2 τιμές να είναι σχετικά κοντά. Έτσι, δεν μπορούμε να χρησιμοποιήσουμε τις μεθόδους που γνωρίζουμε αφού οι μεταβλητές είναι εξαρτημένες, αλλά μπορούμε να τρέξουμε την εντολή t.test στην R, με τη διαφορά ότι κάνουμε paired t-test, αφού θέτουμε paired= TRUE. Επίσης, θέτουμε την εναλλακτική ως μ>=0 και τη μηδενική υπόθεση το συνεργείο να μην υπερεκτιμά τις ζημιές. Παρακάτω έχουμε παραθέσει

το στιγμιότυπο από την R, όπου j εμπεριέχει τις τιμές του συνεργείου και h τις τιμές του εμπειρογνώμονα:

```
> t.test(j, h, paired = TRUE,alternative='greater')

        Paired t-test

data:  j and h
t = 2.9132, df = 9, p-value = 0.008611
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 42.63653      Inf
sample estimates:
mean of the differences
                    115
```

Έχοντας ορίσει σαν H0: μ=0 και Ηα: μ>=0, η μέση τιμή των διαφορών τιμή συνεργείου πλην τιμή εμπειρογνώμονα είναι 115 και το pvalue είναι 0.008611. Όπως αντιλαμβανόμαστε, είναι πολύ μικρό αφού ακόμα και α= 1% να θέταμε, θα ήταν μεγαλύτερο από το pvalue. Αυτό μας οδηγεί στο ασφαλές συμπέρασμα να απορρίψουμε τη μηδενική υπόθεση. Άρα η περίπτωση να μην υπερεκτιμά το συνεργείο τις ζημιές απορρίπτεται.

8^{Ο} Ερώτημα

Α) Αρχικά κάνουμε import τα δεδομένα στην R. Έπειτα, με την ίδια λογική με το 5^{ο} ερώτημα, θέτουμε με την εντολή «d[94,4]<-NA» -όπου d ο πίνακας με τα δεδομένα- το κελί σε αυτή τη θέση ίσο με null, για να μην συμπεριληφθεί στους υπολογισμούς, καθώς περιέχει μία τιμή που είναι τυπογραφικό λάθος. Ύστερα, ελέγχουμε τα δεδομένα μας προτού χρησιμοποιήσουμε οποιαδήποτε μέθοδο συμπερασματολογίας. Αφού τα δεδομένα προέρχονται από τυχαία δειγματοληψία, το n> 15 και δεν έχουμε κάποιο outlier καθώς αγνοούμε την τιμή που ήταν τυπογραφικό λάθος, καταλήγουμε ότι τα δεδομένα είναι κατάλληλα και μπορούμε να χρησιμοποιήσουμε τις ανάλογες μεθόδους συμπερασματολογίας. Έπειτα επιλέγουμε να λύσουμε το ερώτημα με κλήση της t.test όπως φαίνεται παρακάτω:

```
> t.test(d$height[d$sex=="M"], d$height[d$sex=="F"])

        Welch Two Sample t-test

data:  d$height[d$sex == "M"] and d$height[d$sex == "F"]
t = 8.9954, df = 65.471, p-value = 4.745e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.09882401 0.15521810
sample estimates:
mean of x mean of y
 1.793600  1.666579
```

Επομένως, σύμφωνα με την R το ζητούμενο διάστημα είναι 0.09882401 0.15521810.

Β) Θα κάνουμε two sample t test όπως φαίνεται παρακάτω:

```
> t.test(d$prob[d$sex=="M"], d$prob[d$sex=="F"],alternative = 'greater')

        Welch Two Sample t-test

data:  d$prob[d$sex == "M"] and d$prob[d$sex == "F"]
t = -1.1841, df = 71.526, p-value = 0.8798
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.239088       Inf
sample estimates:
mean of x mean of y
 6.323529  6.838235
```

Αφού έχουμε 95% διάστημα εμπιστοσύνης, έχουμε 5% επίπεδο σημαντικότητας. Επομένως ο έλεγχος που κάνουμε, είναι κατάλληλος. . Η μέση τιμή των κοριτσιών, είναι μεγαλύτερη από αυτή των αγοριών, άρα η απάντηση στο ερώτημα είναι, όχι δεν επιτυγχάνουν μεγαλύτερο μέσο βαθμό. Σε αυτό βοηθάει και το α που είναι 0.05 -αφού έχουμε 5% επίπεδο σημαντικότητας- και είναι φανερά μικρότερο από το pvalue, που έχει πολύ μεγάλη τιμή ίση με 0.8798 (και δικαίως, αφού τα κορίτσια φαίνεται ότι έχουν μεγαλύτερους βαθμούς και εμείς σύμφωνα με την εκφώνηση, ορίσαμε ως Ηα την περίπτωση να έχουν τα αγόρια μεγαλύτερο βαθμό, γι'αυτό και η σύγκριση τείνει προς την Η0 η οποία είναι πιο κοντά στα δεδομένα) άρα, οδηγούμαστε στην αποτυχία απόρριψης της μηδενικής υπόθεσης.

Γ) Κάνουμε πάλι ένα two sample t test όπως φαίνεται παρακάτω:

```
> t.test(d$prob, d$math)

        Welch Two Sample t-test

data:  d$prob and d$math
t = -0.60705, df = 205.15, p-value = 0.5445
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7351971  0.3890432
sample estimates:
mean of x mean of y
 6.471154  6.644231
```

Έχουμε θέσει ως Η0: μ=μ0 την υπόθεση, ότι είναι ίδιες οι τιμές στα 2 μαθήματα ενώ ως Ηα: μ!=0 την υπόθεση να διαφέρουν. Το pvalue αγγίζει το 55% και προφανώς δεν μπορεί να απορριφθεί η μηδενική υπόθεση. Η μέση τιμή των πιθανοτήτων, είναι λίγο μικρότερη από τη μέση τιμή των μαθηματικών, χωρίς ωστόσο να έχουν μεγάλη απόκλιση. Επομένως με τα δεδομένα που έχουμε, καταλήγουμε ότι επειδή δεν απορρίψαμε την περίπτωση να είναι ίδιες οι τιμές, δεν σημαίνει ότι είναι ίδιες, απλά παρατηρούμε μία μικρή σχετικά απόκλιση.